Machine learning-guided prediction of OleA thiolase activity and substrate scope

Serina L. Robinson,[1,2,3] Megan D. Smith,[2,3] Jack E. Richman,[3] Kelly G. Aukema,[3] Lawrence P. Wackett[3]

[1]Graduate Program in Bioinformatics and Computational Biology, University of Minnesota 111 S. Broadway, Suite 300, Rochester, MN, 55904, USA; [2]Graduate Program in Microbiology, Immunology, and Cancer Biology, University of Minnesota, 689 23rd Ave SE, Minneapolis, MN, 55455, USA; [3]BioTechnology Institute, University of Minnesota, 1479 Gortner Avenue, Saint Paul, MN, 55108, USA

Main text figures

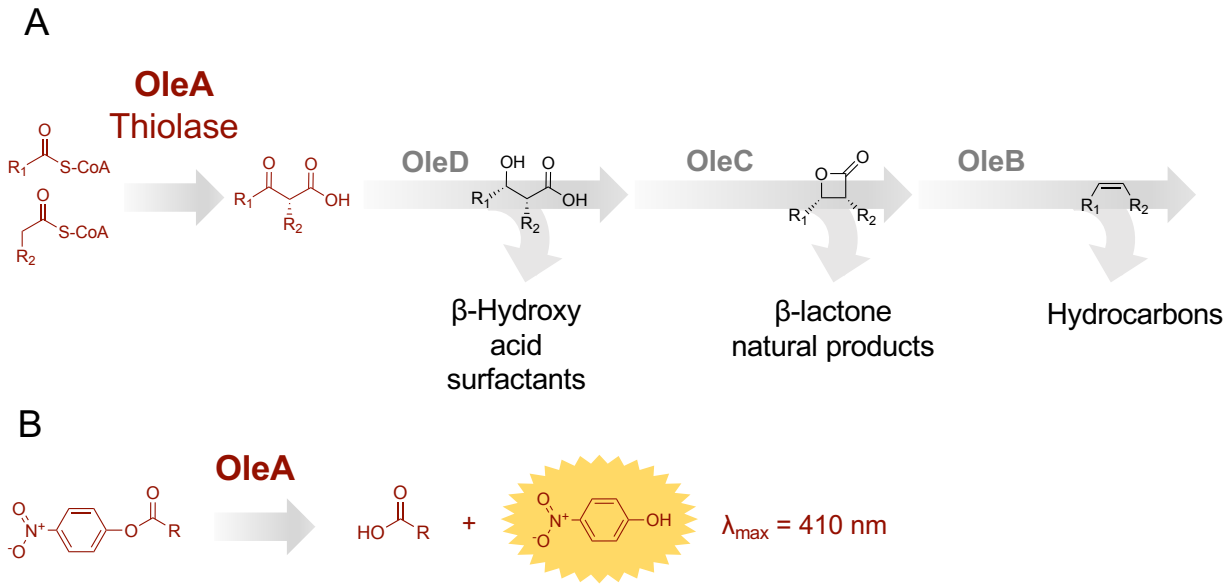
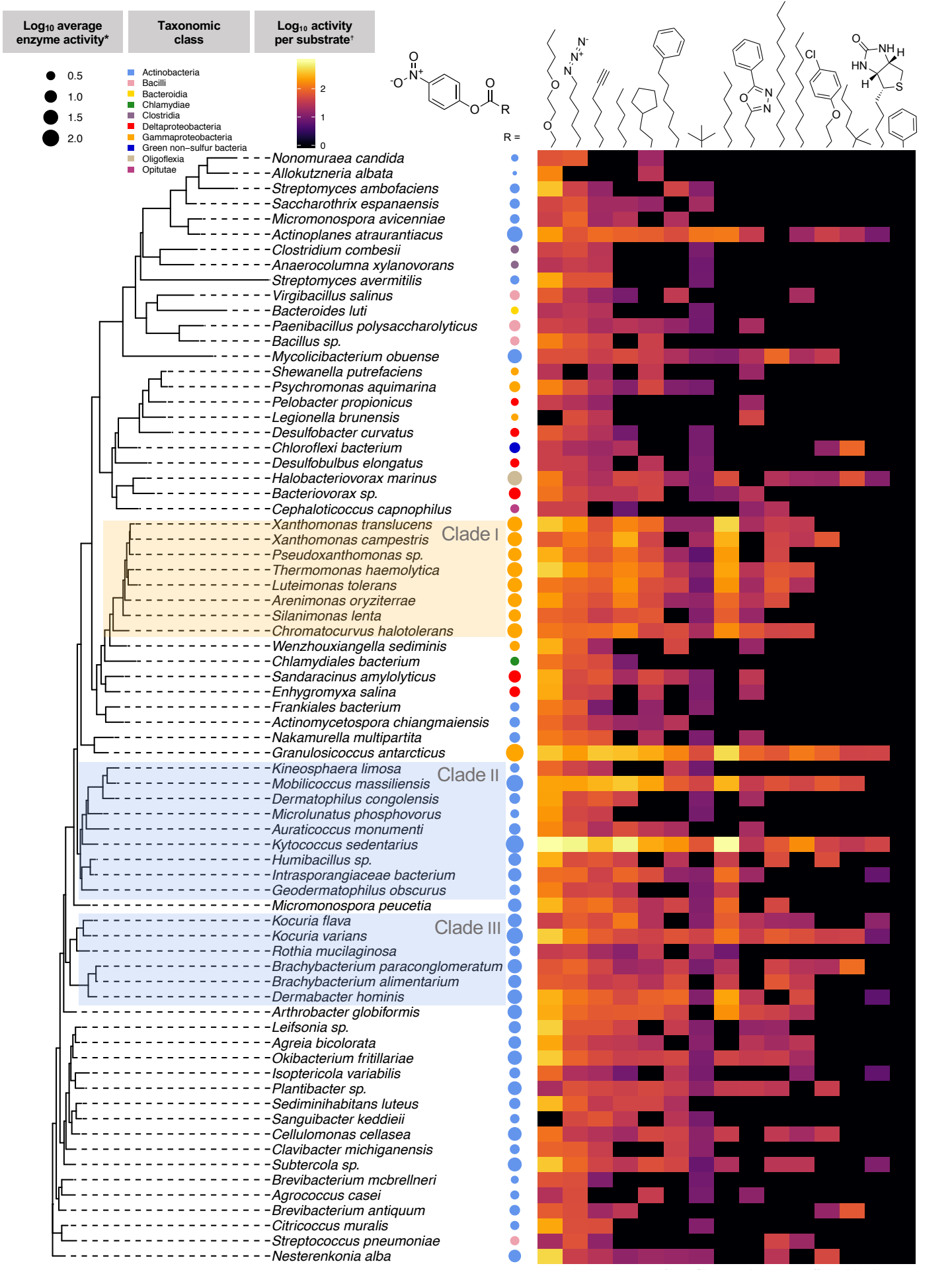


Figure 1. Thiolase enzymes in the OleA family catalyze a head-to-head Claisen condensation of two acyl-CoA substrates (maroon) as the first committed step in production of value-added metabolites such as surfactants, pharmaceuticals and hydrocarbons. R_1, R_2 in native OleA pathways: C_8 – C_{16}. (B) OleA reacts with various *para*-nitrophenyl esters to produce the corresponding carboxylic acids and *para*-nitrophenol chromophore, providing a rapid readout for enzyme activity.


Log10 average enzyme activity*
Taxonomic class
Log10 activity per substrate†
0.5
1.0
1.5
2.0
Actinobacteria
Bacilli
Bacteroidia
Chlamydiae
Clostridia
Deltaproteobacteria
Gammaproteobacteria
Green non−sulfur bacteria
Oligoflexia
Opitutae
2
1
0
R =
Nonomuraea candida
Allokutzneria albata
Streptomyces ambofaciens
Saccharothrix espanaensis
Micromonospora avicenniae
Actinoplanes atraurantiacus
Clostridium combesii
Anaerocolumna xylanovorans
Streptomyces avermitilis
Virgibacillus salinus
Bacteroides luti
Paenibacillus polysaccharolyticus
Bacillus sp.
Mycolicibacterium obuense
Shewanella putrefaciens
Psychromonas aquimarina
Pelobacter propionicus
Legionella brunensis
Desulfobacter curvatus
Chloroflexi bacterium
Desulfobulbus elongatus
Halobacteriovorax marinus
Bacteriovorax sp.
Cephaloticoccus capnophilus
Xanthomonas translucens
Xanthomonas campestris
Pseudoxanthomonas sp.
Thermomonas haemolytica
Luteimonas tolerans
Arenimonas oryziterrae
Silanimonas lenta
Chromatocurvus halotolerans
Clade I
Wenzhouxiangella sediminis
Chlamydiales bacterium
Sandaracinus amylolyticus
Enhygromyxa salina
Frankiales bacterium
Actinomycetospora chiangmaiensis
Nakamurella multipartita
Granulosicoccus antarcticus
Kineosphaera limosa
Mobilicoccus massiliensis
Dermatophilus congolensis
Microlunatus phosphovorus
Auraticoccus monumenti
Kytococcus sedentarius
Humibacillus sp.
Intrasporangiaceae bacterium
Geodermatophilus obscurus
Clade II
Micromonospora peucetia
Kocuria flava
Kocuria varians
Rothia mucilaginosa
Brachybacterium paraconglomeratum
Brachybacterium alimentarium
Dermabacter hominis
Clade III
Arthrobacter globiformis
Leifsonia sp.
Agreia bicolorata
Okibacterium fritillariae
Isoptericola variabilis
Plantibacter sp.
Sediminihabitans luteus
Sanguibacter keddieii
Cellulomonas cellasea
Clavibacter michiganensis
Subtercola sp.
Brevibacterium mcbrellneri
Agrococcus casei
Brevibacterium antiquum
Citricoccus muralis
Streptococcus pneumoniae
Nesterenkonia alba

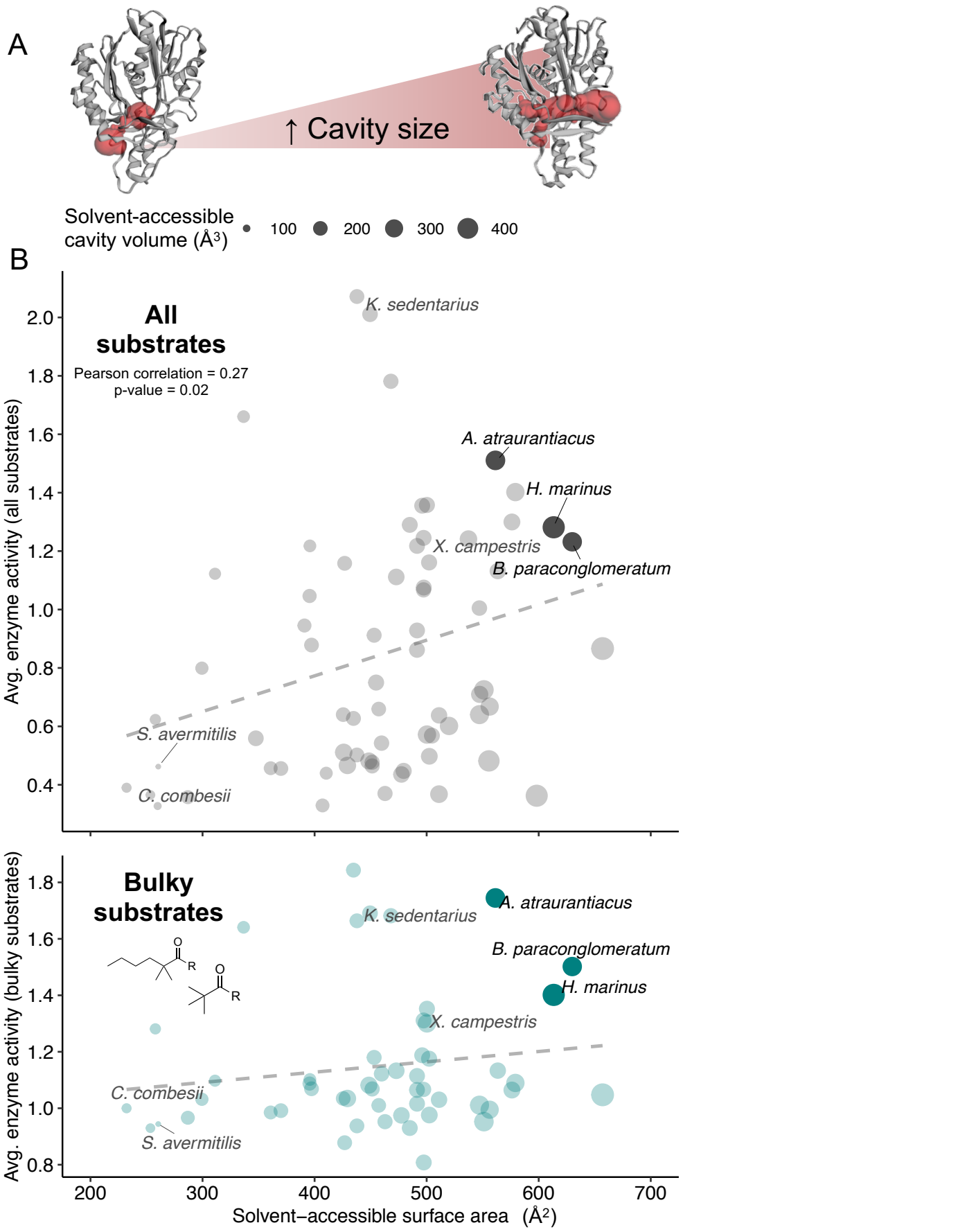


Figure 3. (A) Predicted solvent-accessible cavity volumes of the 73 OleA enzymes analyzed in this study ranged from 95.3 to 478.9 Å³ (B) Relationship between the solvent-accessible surface area with the average enzymatic activity across all substrates (black) and 'bulky' substrates (teal) with tertiary α-carbons (*p*NP trimethylacetate and *p*NP dimethyl hexanoate). *A. atraurantiacus, B. paraconglomeratum*, and *H. marinus* enzymes have among the highest calculated cavity volumes and the highest preferences for substrates with tertiary α-carbons.

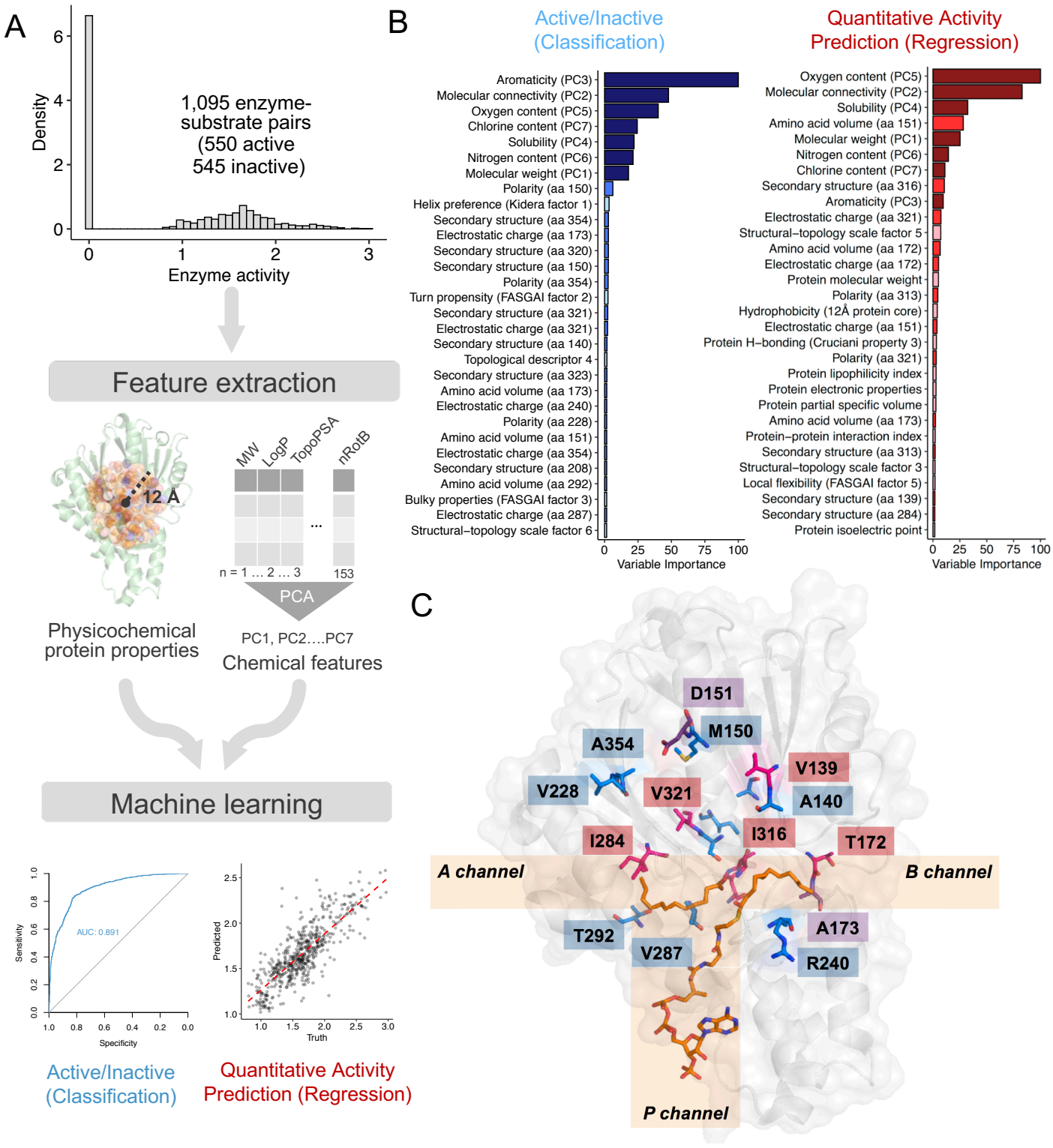


Figure 4. (A) Machine learning workflow (B) Variable importance scores for classification and regression models (C) Important residues mapped onto the *X. campestris* 4KU5 structure with fatty acid substrates bound. Residue colors correspond to variable importance in the classification model (blue), regression model (red), or both models (purple).

Table 1. Machine learning classification results. See Fig S3 for extended results.

Machine learning algorithm	Training classification accuracy	Testing classification accuracy	Testing 95% confidence interval
Random Forest	0.826	0.839	(0.789, 0.880)
Feedforward Neural Network	0.732	0.777	(0.722, 0.826)
Naïve Bayes	0.586	0.645	(0.585, 0.701)

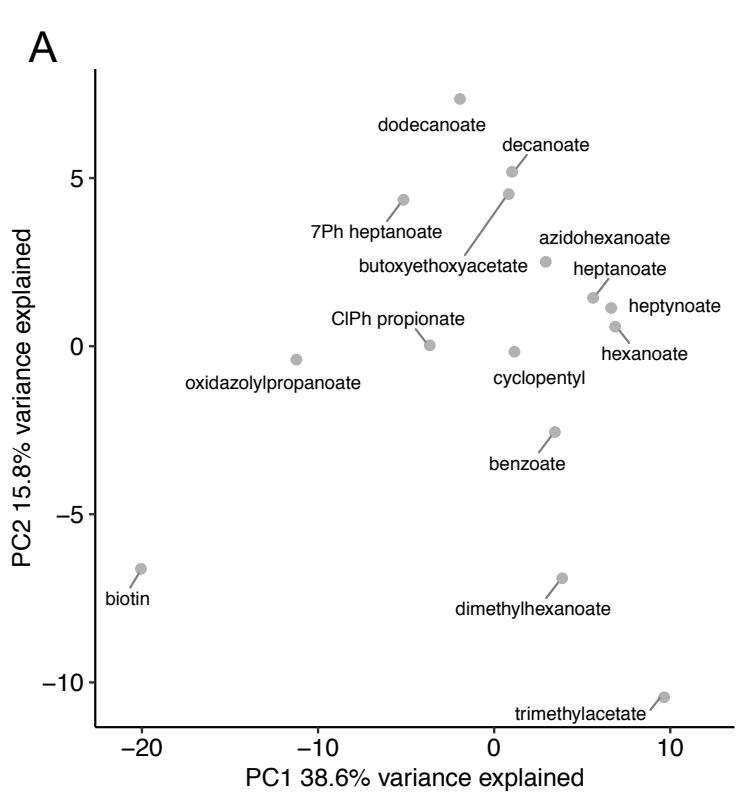
Table 2. Machine learning regression results. See Fig S3 for extended results.

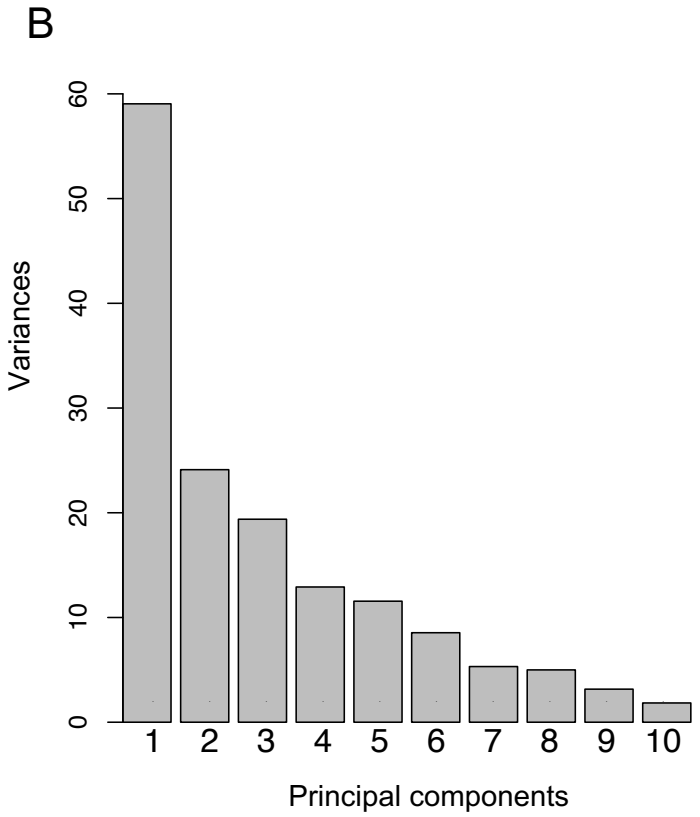
Machine learning algorithm	Training RMSE	Training R^2	Testing RMSE	Testing R^2
Random Forest	0.254	0.625	0.219	0.745
Multivariate adaptive regression splines	0.276	0.557	0.252	0.642
Elastic Net	0.275	0.549	0.278	0.564

Supplemental Figures

Table S1. Accession numbers, organisms, average activity across all substrates, and total number of substrates accepted (out of 15). †Enzyme activity is measured in the $\log_{10}$ nmol *p*NP produced over the course of one hour by 200 µL *E. coli* BL21 culture with an OD of 1.0 heterologously expressing OleA averaged across three independent biological replicates and across all substrates.

NCBI Accession	Organism	Average enzyme activity†	Total number of substrates accepted
WP_012802190	*Kytococcus sedentarius*	2.072	14
WP_088919872	*Granulosicoccus antarcticus*	2.011	14
WP_040156198	*Mobilicoccus massiliensis*	1.781	13
WP_068469535	*Kocuria varians strain G6*	1.66	14
WP_097326105	*Actinoplanes atraurantiacus*	1.511	13
WP_114959563	*Thermomonas haemolytica*	1.402	11
WP_117315975	*Chromatocurvus halotolerans*	1.358	11
WP_003480168	*Xanthomonas translucens pv. graminis*	1.355	11
WP_101686474	*Dermabacter hominis*	1.3	11
WP_076585679	*Luteimonas tolerans*	1.29	11
WP_096909276	*Halobacteriovorax marinus strain BE01*	1.282	13
NP_635607	*Xanthomonas campestris*	1.245	10
WP_003805730	*Arthrobacter globiformis*	1.241	10
WP_126985334	*Brachybacterium paraconglomeratum*	1.232	12
WP_079726385	*Okibacterium fritillariae*	1.218	11
WP_082133371	*Mycolicibacterium obuense*	1.218	12
WP_022969495	*Arenimonas oryziterrae*	1.161	10
WP_104244666	*Subtercola sp. Z020*	1.158	11
WP_086472704	*Plantibacter sp. VKM Ac-1784*	1.132	11
WP_058859095	*Kocuria flava*	1.122	11
WP_034630949	*Cellulomonas cellasea*	1.111	11
WP_130552392	*Pseudoxanthomonas sp. NML171200*	1.075	9
WP_091632192	*Micromonospora peucetia*	1.067	9
WP_026862459	*Intrasporangiaceae bacterium URHB0013*	1.046	10
WP_096166078	*Brachybacterium alimentarium*	1.005	9
WP_070193933	*Humibacillus sp. DSM 29435*	0.945	8
WP_044441590	*Agreia bicolorata*	0.928	9
WP_033420292	*Nesterenkonia alba*	0.912	9
WP_028771085	*Silanimonas lenta*	0.878	8
WP_053234550	*Sandaracinus amylolyticus*	0.866	8
WP_056164922	*Leifsonia sp. Leaf325*	0.862	8
WP_021267235	*Bacteriovorax sp. BAL6 X*	0.799	8
WP_090596165	*Auraticoccus monumenti*	0.75	7
WP_090922125	*Paenibacillus polysaccharolyticus*	0.725	8
WP_100422947	*Sediminihabitans luteus*	0.709	6
WP_028865527	*Psychromonas aquimarina*	0.678	7
WP_013837355	*Isoptericola variabilis*	0.668	8
WP_028327376	*Dermatophilus congolensis*	0.659	6
OGN97459	*Chloroflexi bacterium RBG 13 51 36*	0.648	7
WP_015749354	*Nakamurella multipartita*	0.64	6
WP_012948569	*Geodermatophilus obscurus*	0.64	6
WP_106091803	*Enhygromyxa salina strain SWB007*	0.638	6
WP_101620581	*Brevibacterium antiquum*	0.627	6
WP_005508112	*Rothia mucilaginosa*	0.623	7
WP_026204600	*Actinomycetospora chiangmaiensis*	0.601	6
WP_015101236	*Saccharothrix espanaensis*	0.572	6
WP_116651321	*Wenzhouxiangella sediminis*	0.568	5
WP_092493676	*Virgibacillus salinus*	0.559	6
WP_015490789	*Clavibacter michiganensis*	0.543	5
WP_053141838	*Streptomyces ambofaciens*	0.542	5
WP_076469753	*Micromonospora avicenniae*	0.511	5
WP_006592558	*Kineosphaera limosa*	0.502	5
WP_012868615	*Sanguibacter keddieii*	0.498	5
WP_119951226	*Frankiales bacterium YIM 75000*	0.482	5
WP_095301808	*Bacillus sp. 7586-K*	0.482	4
CVN04163	*Streptococcus pneumoniae*	0.477	5
WP_051554118	*Desulfobulbus elongatus*	0.466	5
WP_115933252	*Citricoccus muralis*	0.465	4
WP_010983715	*Streptomyces avermitilis*	0.462	4
WP_020589018	*Desulfobacter curvatus*	0.457	5
WP_086990725	*Agrococcus casei*	0.456	5
WP_068712879	*Cephaloticoccus capnophilus*	0.448	5
WP_088206054	*Chlamydiales bacterium SCGC AG-110-P3*	0.44	4
WP_013865923	*Microlunatus phosphovorus*	0.435	4
WP_099839552	*Clostridium combesii*	0.39	4
WP_005885358	*Brevibacterium mcbrellneri*	0.37	4
WP_061782733	*Shewanella putrefaciens*	0.368	4
WP_084558798	*Anaerocolumna xylanovorans*	0.365	4
ABL00697	*Pelobacter propionicus DSM 2379*	0.362	4
WP_073399371	*Bacteroides luti*	0.358	4
WP_063765283	*Nonomuraea candida*	0.329	3
WP_058441655	*Legionella brunensis*	0.327	3
WP_030433686	*Allokutzneria albata*	0.243	2





C

PC1: Molecular weight index		PC2: Molecular connectivity index		PC3: Aromaticity index		PC4: Solubility index		PC5: Oxygen content index		PC6: Nitrogen content index		PC7: Chlorine content index	
PC1	PC1 loading	PC2	PC2 loading	PC3	PC3 loading	PC4	PC4 loading	PC5	PC5 loading	PC6	PC6 loading	PC7	PC7 loading
ATSm2	0.12733898	VC.5	0.17825627	HybRatio	0.20968749	MLogP	0.17154508	MDEO.22	0.25897898	khs.dsN	0.25341943	ATSc5	0.31651235
SP.6	0.12691363	VC.3	0.17746436	nAromBond	0.19391724	BCUTw.1l	0.15862225	khs.ssO	0.25102431	MDEN.12	0.25341943	khs.sCl	0.28076896
SP.4	0.12645294	SC.3	0.17703505	naAromAtom	0.19391724	khs.sCH3	0.15763772	MDEO.12	0.24672596	MDEN.13	0.25341943	Cl	0.28076896
ATSp2	0.12645069	SPC.4	0.17585691	AROMATIC	0.19284941	BCUTp.1l	0.14998059	C1SP3	0.24131202	Ncharges	0.25341943	khs.aaN	0.23438633
ATSp1	0.12620083	Kier3	0.17310593	khs.aasC	0.1915478	khs.dsN	0.14679924	BCUTc.1l	0.2399008	BCUTp.1l	0.24831762	khs.aaO	0.23438633
Zagreb	0.12584608	Kier2	0.17192908	C2SP2	0.18874054	MDEN.12	0.14679924	ROR	0.23947382	MDEN.22	0.24652643	C1SP2	0.23438633
SP.7	0.12574425	SC.5	0.16981795	khs.aaCH	0.18672388	MDEN.13	0.14679924	O	0.23482976	N	0.23141734	BCUTw.1h	0.2307478
ATSm3	0.12517459	VPC.4	0.15951925	ALogp2	0.18128494	Ncharges	0.14679924	WTPT.4	0.23468926	WTPT.5	0.2209668	C1SP1	0.20580706
ATSm5	0.12472649	nRotB	0.15864429	MolIP	0.17771233	XLogP	0.1453058	ATSc1	0.2322058	MDEN.23	0.20521987	C2SP1	0.20580706
SP.3	0.12467235	MDEC.22	0.15689146	ALogP	0.15681012	MDEC.24	0.13925752	ATSc3	0.22603973	ALogP	0.12168195	khs.tCH	0.20580706
SP.5	0.12391214	khs.ssCH2	0.15017	C2SP3	0.15518055	MDEC.12	0.1373312	ATSc4	0.22361334	BCUTc.1l	0.12123668	khs.tsC	0.20580706
nB	0.12362396	VC.4	0.14863777	C3SP2	0.15143774	N	0.13637284	ATSc2	0.20017451	C1SP2	0.12046929	RCCH	0.20580706
ATSp5	0.12360823	VC.6	0.14863777	bpol	0.14974329	C4SP3	0.13603753	XLogP	0.16485005	khs.aaN	0.12046929	ATSc4	0.13422367
SP.2	0.12335411	SC.6	0.14863777	nAtomLAC	0.1472842	khs.ssssC	0.13603753	nHBAcc	0.16004107	khs.aaO	0.12046929	topoShape	0.13137093
MW1	0.12311343	SC.4	0.14863777	khs.ssCH2	0.14166215	MDEC.34	0.13603753	MLogP	0.15145252	khs.ssO	0.11856319	Petitjean Number	0.12885204
MW	0.12311343	MDEC.14	0.14844248	FMF	0.13248809	C	0.13053089	WTPT.3	0.13982385	BCUTc.1h	0.10954866	Ncharges	0.11752702
ATSp3	0.12307697	MDEC.11	0.14756262	nAtomLC	0.12789728	WTPT.5	0.129784	BCUTc.1h	0.13594845	BCUTw.1h	0.10936667	khs.dsN	0.11752702
ATSp4	0.12298691	C4SP3	0.14603078	fragC	0.12775236	MDEC.13	0.12905554	ATSc5	0.12833043	BCUTw.1l	0.10261149	MDEN.12	0.11752702
VAdjMat	0.12186202	khs.ssssC	0.14603078	khs.aaN	0.12639394	VC.4	0.12688135	MDEC.22	0.11822117	nHBAcc	0.10032192	MDEN.13	0.11752702
ATSm4	0.1209683	MDEC.34	0.14603078	khs.aaO	0.12639394	VC.6	0.12688135	C2SP3	0.11215285	C1SP1	0.09628498	ATSm1	0.11675954
MW2	0.12088089	nAtomLC	0.14480924	C1SP2	0.12639394	SC.6	0.12688135	FMF	0.09340499	C2SP1	0.09628498	BCUTp.1l	0.11613374
WTPT.1	0.1207875	SPC.6	0.13705444	nAtom	0.11438864	SC.4	0.12688135	C3SP2	0.09094711	khs.tCH	0.09628498	C2SP2	0.11360216
WPOL	0.12071096	LipinskiFailures	0.13648816	MDEC.12	0.11134	LipinskiFailures	0.1251014	TopoPSA	0.08888915	khs.tsC	0.09628498	BCUTc.1l	0.11056927
ATSm1	0.11818479	SPC.5	0.13300536	MDEC.23	0.10565009	MDEC.14	0.12314095	C	0.07962149	RCCH	0.09628498	SCH.5	0.10960208
SP.1	0.11737146	MDEC.13	0.1325928	nRotB	0.10278535	MDEC.11	0.11808087	MDEC.23	0.07853418	WTPT.3	0.09027589	BCUTw.1l	0.10614501
SCH.7	0.11724575	BCUTp.1h	0.13214258	VP.2	0.1020575	fragC	0.11446085	khs.aaCH	0.07085603	khs.sCl	0.09004652	MDEC.12	0.10015262
SCH.6	0.11669521	VPC.5	0.12064741	VP.7	0.10171919	nAtom	0.11386639	nAtomLAC	0.06962266	Cl	0.09004652	ALogp2	0.09479929
VP.5	0.11580149	MDEC.24	0.12014353	VP.6	0.10017882	SC.5	0.11241815	nAtomLC	0.06807243	O	0.08103433	MDEN.23	0.08970466
AMR	0.11566765	XLogP	0.11944323	WTPT.2	0.09899299	bpol	0.10942052	C3SP3	0.06252011	VCH.5	0.07999815	MolIP	0.0836205

Figure S2. (A) Dimensionality reduction of molecular descriptors of 15 *p*NP substrates using principal component analysis. Substrates are plotted by their first two principal components that cumulatively explain up to 54.4% of the variance between compounds (B) Screeplot of principal components reveals 'elbow' after seventh principal component. (C) Absolute values of loadings of 149 molecular descriptors into 7 principal components roughly corresponding to molecular weight, molecular connectivity, aromaticity, solubility, and O, N and Cl content.

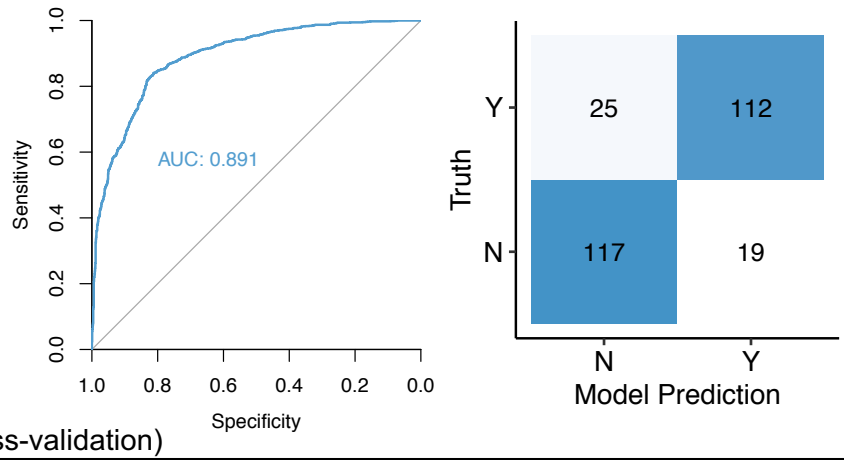
A Classification Results

Model comparison and hyperparameter tuning (10-fold cross-validation)

Machine learning algorithm	Training classification accuracy	Training Cohen's kappa	Testing classification accuracy	Testing 95% confidence interval	Best model hyperparameters
Random Forest	0.826	0.651	0.839	(0.789, 0.880)	mtry = 334, splitrule = extratrees, min.node.size = 1
Feedforward Neural Network	0.732	0.170	0.777	(0.722, 0.826)	size = 3, decay = 0.5
Naïve Bayes	0.586	0.645	0.645	(0.585, 0.701)	fL = 0, adjust = 1, useKernel = F

Average classification results 10 different train-test splits	
Training set accuracy	0.819 ± 0.009
Testing set accuracy	0.808 ± 0.017

Hold-out test set receiver-operator characteristic curve and confusion matrix



B Regression Results

Model comparison and tuning (10-fold cross-validation)

Machine learning algorithm	Training root-mean-square error (RMSE)	Training R^2	Training mean absolute error	Testing RMSE	Testing R^2	Best model hyperparameters
Random Forest	0.254	0.625	0.198	0.219	0.745	mtry = 168, splitrule = variance, min.node.size = 5
Multivariate adaptive regression splines	0.276	0.557	0.217	0.252	0.642	nprune = 20, degree = 1
Elastic Net	0.275	0.549	0.214	0.278	0.564	alpha = 1 (Lasso) lambda = 0.0037

Average regression results 10 different training-test splits	
Training set RMSE	0.249 ± 0.007
Testing set RMSE	0.245 ± 0.019

Hold-out test set regression results and residual plot

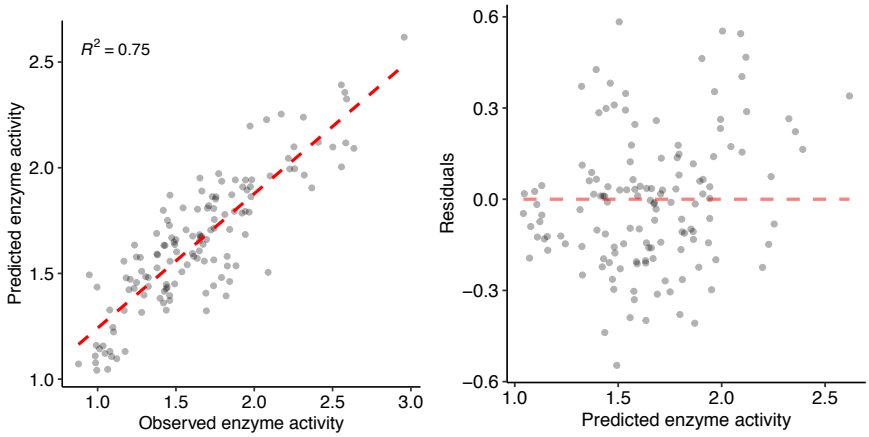
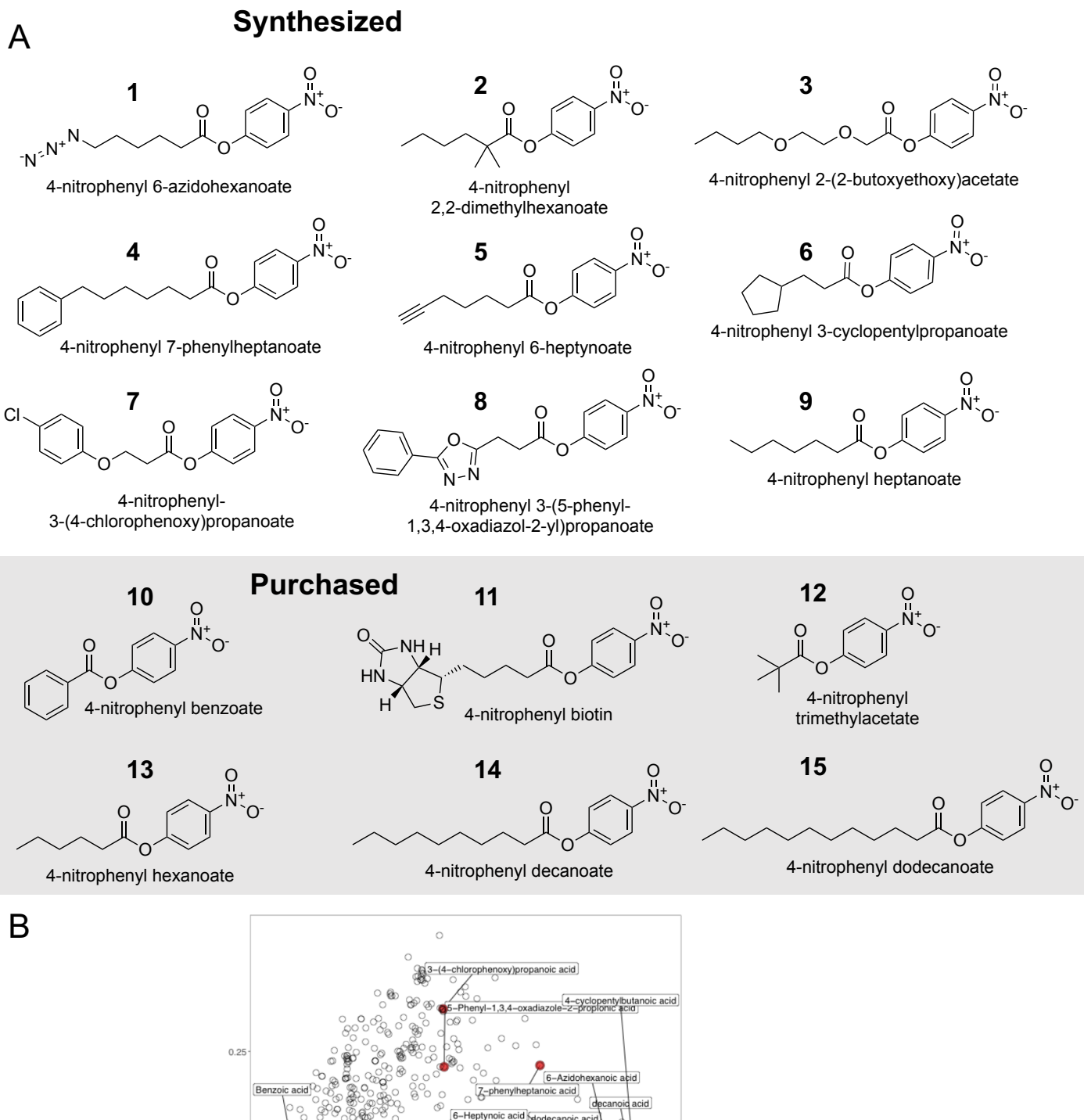


Figure S3. (A) Machine learning classification results and (B) regression results.



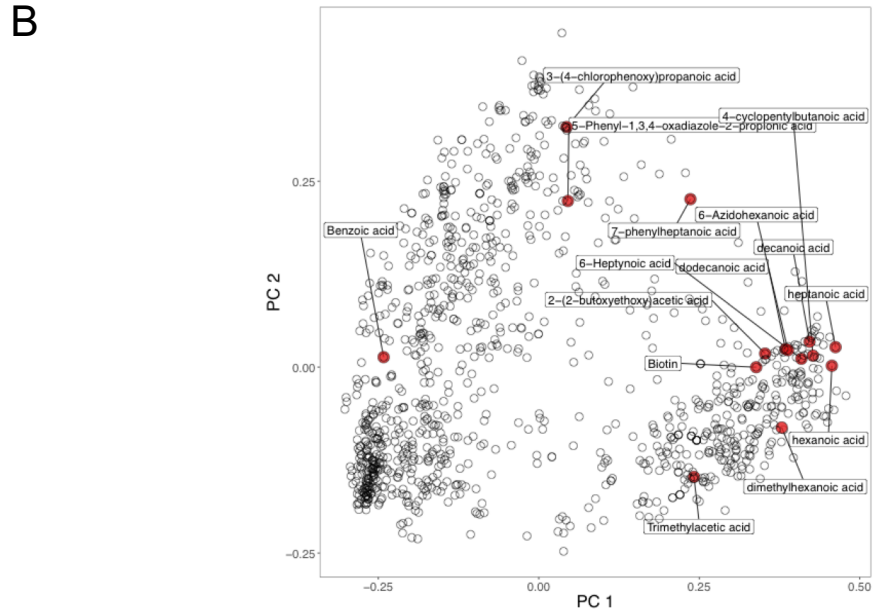
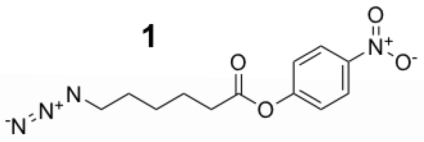
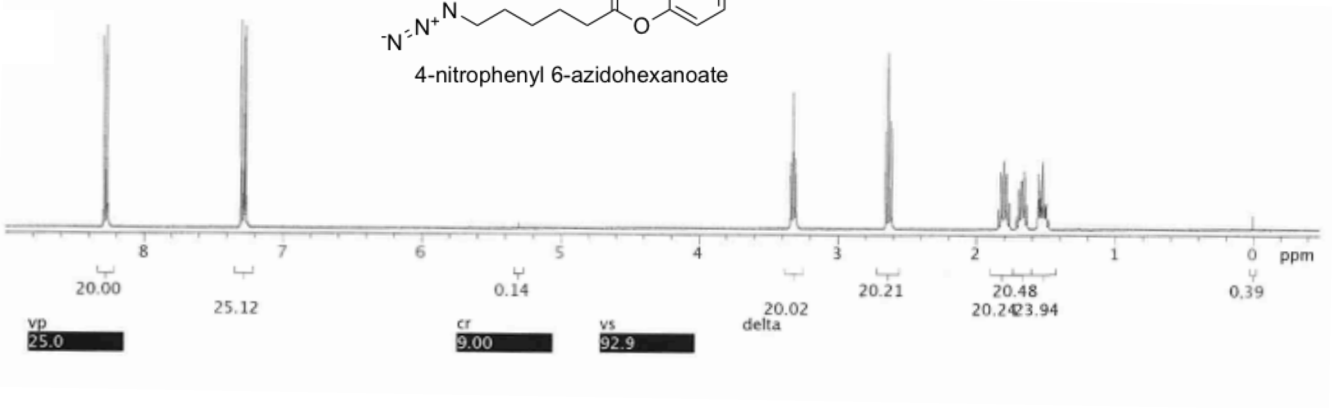


Figure S4. (1) Chemical structures of *p*NP esters synthesized and purchased in this study. (B) Tanimoto clustering of 15 *p*NP substrates (maroon) compared to the sequence space of commercially-available carboxylic acid substrates from Sigma-Aldrich.

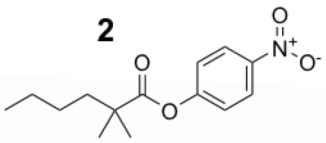
1



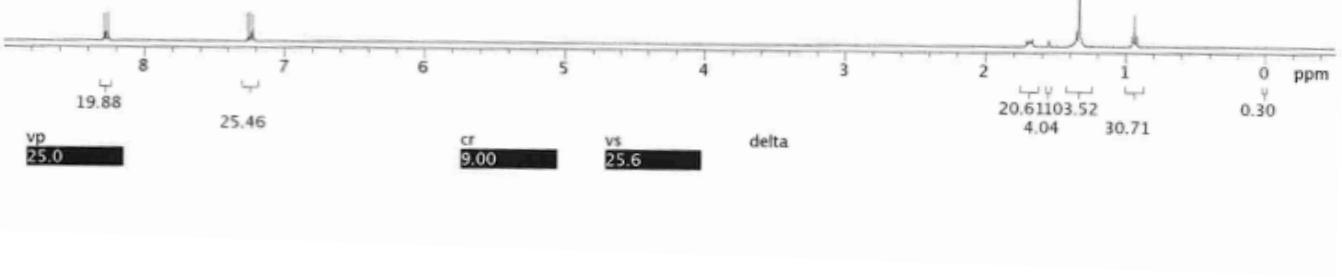
4-nitrophenyl 6-azidohexanoate



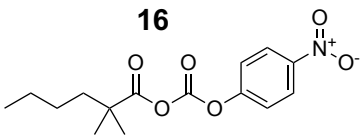
2



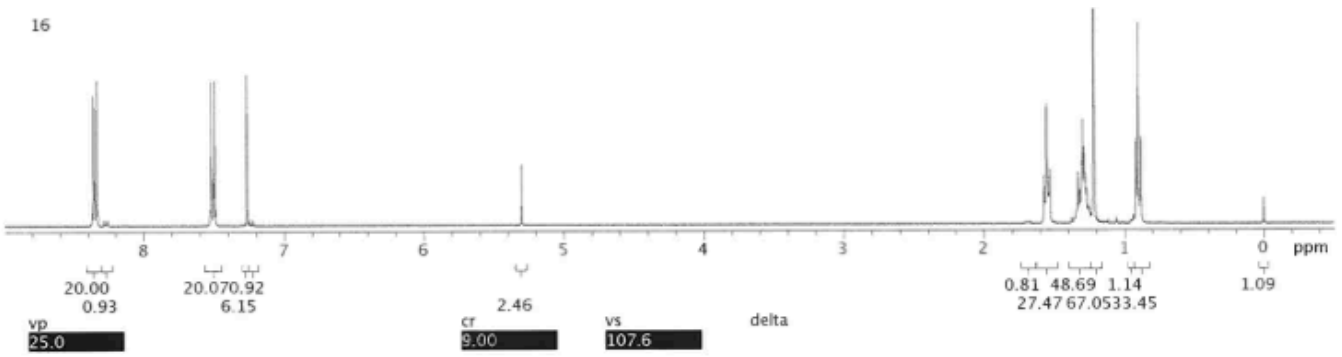
4-nitrophenyl
2,2-dimethylhexanoate



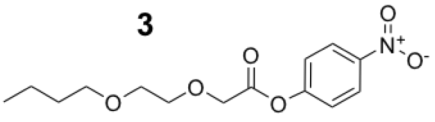
16



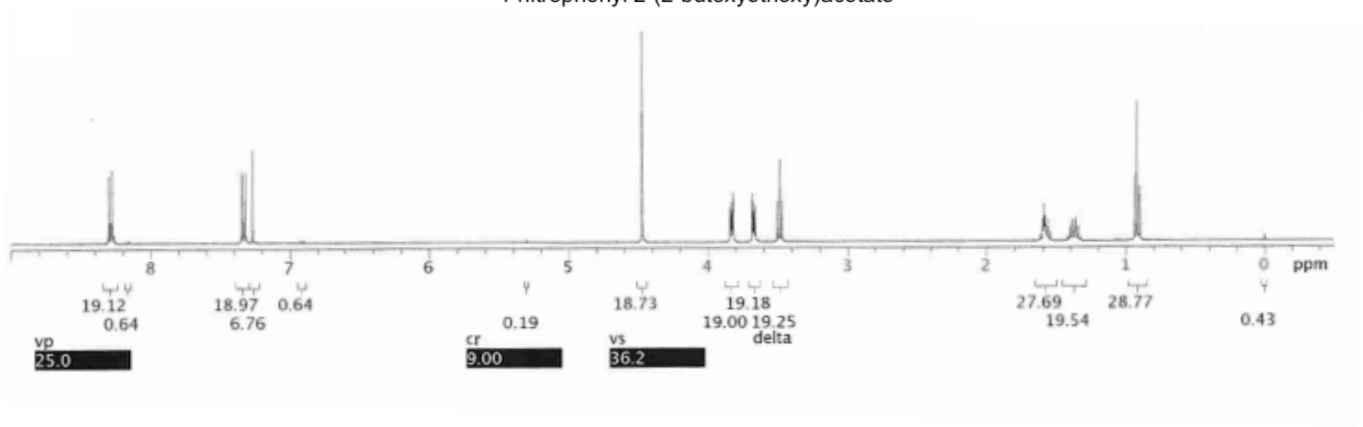
2,2-dimethylhexanoic
(4-nitrophenyl carbonic) anhydride



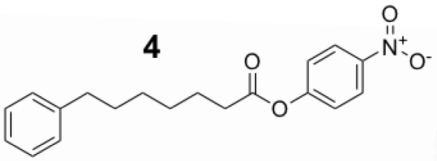
3



4-nitrophenyl 2-(2-butoxyethoxy)acetate

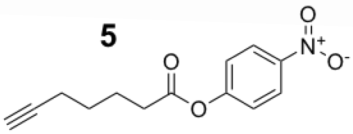


4



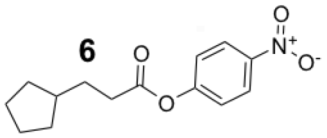
4-nitrophenyl 7-phenylheptanoate

5



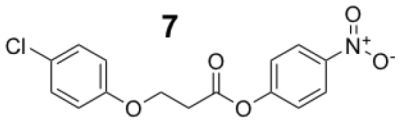
4-nitrophenyl 6-heptynoate

6

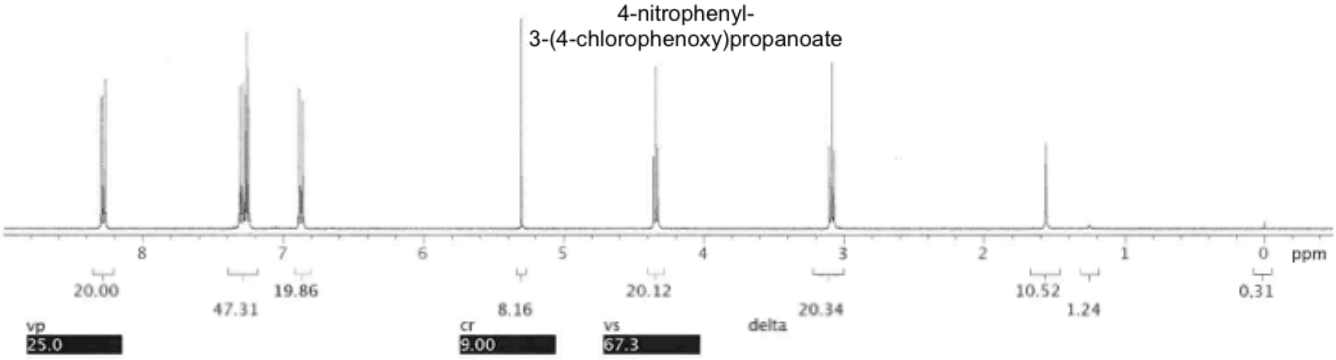


4-nitrophenyl 3-cyclopentylpropanoate

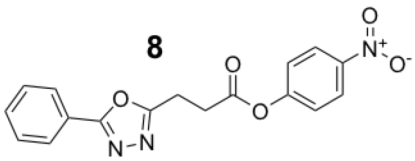
7



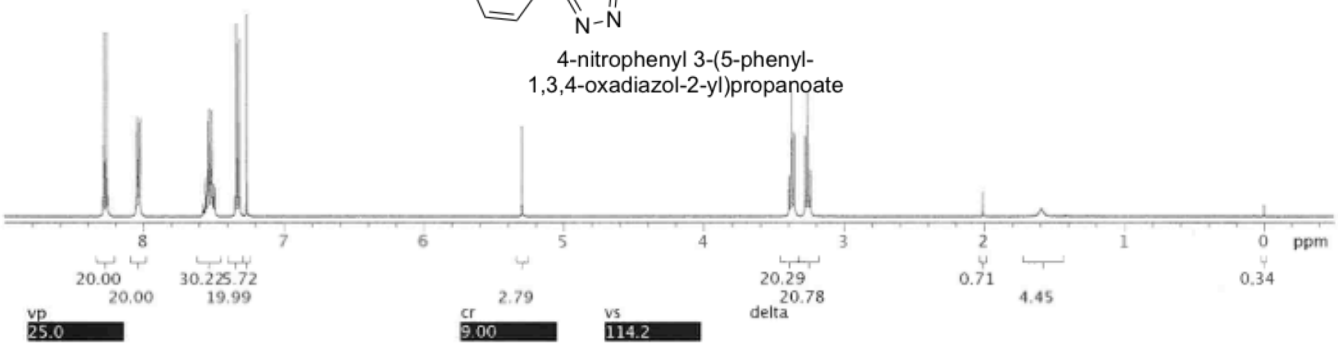
4-nitrophenyl-
3-(4-chlorophenoxy)propanoate



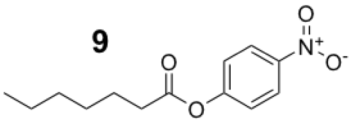
8



4-nitrophenyl 3-(5-phenyl-
1,3,4-oxadiazol-2-yl)propanoate

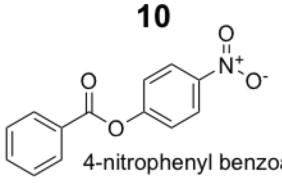


9



4-nitrophenyl heptanoate

10



4-nitrophenyl benzoate